Dear PCI recommender and staff,

We would like to thank the reviewers for their very useful comments. We have now revised the manuscript and updated it on biorxiv. In particular, we added a section addressing the problem of interpretation of parameters of interest in the presence of the r_i's nuisance parameters, implemented the Gamma+neutral model, clarified our model choice strategy, improved the simulation analysis, and added three new supplementary figures. We also rewrote parts of the abstract, introduction and discussion. Below is a point-by-point response to the reviewers' comments. We hope the new version will be considered suitable for recommendation by the PCI Evolutionary Biology.

Best regards

Nicolas Galtier, Marjolaine Rousselle

Reviewer 1

The effective size of a population is the census size of a Wright-Fisher population that would give the same value of some statistic, related to drift. When the Wright-Fisher assumptions are violated, different statistics can imply different values of Ne. Differences between different methods of estimating Ne (e.g. from mean time to pairwise coalescence, vs. total genealogy length) are interesting, because they tell us about violations of the Wright-Fisher assumptions, and because the different Ne values might affect different things (e.g. total neutral diversity vs. current efficacy of purifying selection).

This preprint compares the extent of differences in Ne, estimated in two ways. First, it uses piS, the mean pairwise differences at synonymous sites, where E(piS)=4Nemu. Second, it estimates S=4Nes using the non-synonymous and synonymous SFS. Estimates are obtained for a wide variety of animal species, but focussing on primates vs. Drosophila. Results show that S/min(S) is 1-2 orders of magnitude more variable than piS/min(piS). This is a surprising result, especially if the DFE is identical between species.

The authors relate their finding to Lewontin's paradox: the finding that genetic diversity between species varies much less than would be expected under an equilibrium neutral model, assuming that effective population sizes are proportional to current census sizes. The authors implicate non-equilibrium demography, because piS is strongly influenced by past periods of low population size, while piN/pi_S equilibrates more rapidly, and so should be less influenced by past demography. This point is illustrated with simulations.

The preprint presents interesting work on an important topic, but I think some major points need to be clarified and developed before the results can be properly understood.

Thanks much for a positive evaluation and constructive comments.

1- What is being estimated?

This preprint compares two different types of Ne, but I think the authors could be much clearer about what they measure.

1.1 - In the equation E(pi_S)=4Nemu, 2Ne describes the average time to coalescence of two randomly chosen sequences.
The method of Galtier (2016) estimates Ne in two different ways, via the compound parameters theta=4Nemu and S=4Ne*s. Only the second measure is reported, and I could not tell when/whether the two Ne values are expected to differ from each other, and whether the second measure (which is the only one reported) is a valid measure of Ne.

From eqs. 6 and 10 of Galtier 2016 with r1=1, it looks like 4Ne estimated from theta/mu measures the mean length of the terminal branches on the genealogy. With non-equilibrium demography, this could differ from the mean time to pairwise coalescence. The lengths of the internal branches are fit via the nuisance parameters (ri).

Indeed in Eyre-Walker's r_i-based approach, θ multiplies the expected number of SNPs at every frequency. Because r_1 is set to one and the other r_i's are free parameters, the estimate of θ with this method presumably relies heavily on the density of singletons, i.e., SNPs at frequency 1 in the sample. Please note that assuming neutrality, panmixy and mutation-drift equilibrium in a Wright-Fisher population, the density of singletons is indeed an unbiased estimate of θ. We did not, however, use this estimate in the current study. We rather used the more classical π_S, i.e., the average heterozygosity at synonymous positions, as our estimate of θ. This estimate, which is also unbiased under neutrality and panmixy at mutation-drift equilibrium, takes information from all classes all SNPs. We now explicitly clarify this in the Material and Methods section (p8, l233-234).

1.2 - I found the other Ne (estimated from S/4s) much more difficult to interpret. S is estimated from the complete SFS of non-synonymous mutations. But it does not use existing results for the SFS with selection and non-equilibrium demography or linkage effects (e.g. Evans et al 2007 TPB; Good et al 2014 PLoS Genet). Instead, it uses the standard equilibrium formula for the population allele frequency (eq. 4 of Galtier 2016), while also allowing for non-equilibrium effects on the shape of the underlying genealogy, by allowing for arbitrary variation in the r_i.

For this reason, I struggled to understand what these Ne estimates meant, and whether all of the possible variation in the non-synonymous SFS could be modelled via variation in Ne, even assuming that the gamma+method DFE is accurate.

Because the variation in these Ne values is the major result of the paper, I think this needs to be clarified.

Thanks for a thoughtful comment. The existing theory for non-equilibrium demography and linkage effects is not directly applicable, we think, to the kind of analysis we are doing here. The two papers mentioned by the reviewers do not provide closed formulas for the synonymous and non-synonymous SFS, which are required for efficient likelihood calculation. Importantly, if such formulas existed, they would involve additional parameters (e.g. time-variance in N_e, gene density, recombination landscape). Whether there is enough power to estimate these based on the kind of data we analyze here is far from clear (e.g. see Messer & Petrov 2013). For these reasons, we rather relied on Eyre-Walker's idea of using nuisance parameters, the r_i's, in order to account

for any departure from model assumptions that would similarly affect synonymous and non-synonymous mutations – including linked selection and demographic effects.

Which leaves the question, logically asked by the reviewer, of the meaning of N_e in the equations of Eyre-Walker et al (2006) with r_i's. Galtier (2016) calls it "the size of a Wright-Fisher population that would confer the same amount of distortion between non-synonymous and synonymous SFS as observed in current data" (given the DFE) – admittedly a vague definition. We're here touching a limitation of the r_i-based approach: the inclusion of these nuisance parameters somehow blurs the interpretation of parameters of interest, including N_e and S.

We did two things in order to, at least partly, address this issue. First, we used a version of the model in which all r_i's are set to 1 – so, assumed panmixy, demographic equilibrium and no linked selection. The log-likelihood dropped badly, indicating that such a simple model does not fit the data well enough; the r_i's appear needed. Secondly, we plotted the estimated r_i's in primates and fruit flies (new figure S7) and found that they do not differ dramatically from 1, or across groups. This suggests that the inclusion of these nuisance parameters probably does not affect our analysis strongly – although we acknowledge that a deeper analysis of this potential problem would be worthwhile.

We added a section (p15-16) clarifying these aspects, presenting these new analyses, and discussing the merits and caveats of the r_i-based approach. Although imperfect, it is in our opinion the only existing method suitable for our purpose here.

2- Model adequacy

Estimates of S are highly model-dependent, and the authors present some suggestive evidence that the standard gamma model underfits the SFS data. They use a gamma+lethal model and show that the extra parameter has a substantial influence on the estimates.

2.1- This is an interesting result, but I think that the authors might be too quick to assume that the gamma+lethal model solves the problem of model adequacy, and to conclude that there is little difference in the DFE across their data sets.

We don't mean to firmly conclude that there is little difference in the DFE across data sets. For instance p17 we discuss the possibility that the among-species difference in shape we infer when fitting a Gamma DFE might be, at least in part, of biological origin. We now make even clearer that our discussion of the among-species variation in S is conditional on our assumption of a common DFE (p18, l528-529).

2.2- Previous authors have examined the adequacy of the gamma model, with different sorts of data, including Nielsen & Yang (2003, MBE) and Loewe and Charlesworth (2006, Biol Lett). Nielsen and Yang also used a gamma+lethal model (and a normal + lethal model), while Loewe and Charlesworth argued for a lognormal model. Other authors have combined a gamma distribution with a class of purely neutral mutations (Loewe et al 2006, Genetics; Betancourt et al. 2012 Evolution). It would be useful to know if results are robust to using other distributions like this. While Table 2 contains many good robustness analyses, I think more formal work on model adequacy could be presented to make the headline results really convincing.

Galtier (2016) and Rousselle et al. (2018) introduced and analyzed a number of distinct models of the DFE, based on the Gamma distribution, Beta distribution, and various versions of Fisher's Geometric Model, with or without a beneficial component. Here we additionally implemented the reflected Gamma and Gamma+lethal models. These are suitable for the purpose of this study in that their parametrization makes it easy to model a common shape, and only differences in scale, across data sets. This would be less easily achieved with the ScaledBeta or GammaExpo models, for instance, which performed well in Galtier's (2016) model comparison.

The reviewer mentions the lognormal and Gamma+neutral distributions as two additional options – among many one could think of. Regarding the lognormal distribution, one problem highlighted by Welch et al (2008) is the suppression of density near the origin, for $S \to 0$. This might be compensated by adding a class of neutral mutations, but why relying in the first place on a distribution with such an undesired feature appears questionable.

We did implement the Gamma+neutral distribution suggested by the reviewer, with a fixed proportion of neutral non-synonymous mutations among species. Analyzing our primates+drosophila data set, we found that the likelihood was maximal when the proportion of neutral mutations was zero – so, the additional class did not improve the fit. The likelihood was essentially insensitive to the mass of the neutral class for values between 0 and 0.4, and declined substantially when this parameters exceeded 0.4. We did not modify the manuscript based on these suggestions.

2.3- Second, the authors argue that their estimated likelihood surface "suggests that the DFE perhaps does not differ so dramatically between primates and fruit flies", but doesn't this flat likelihood surface suggest instead that the parameters are non-identifiable with data of this kind?

"Non-identifiable" is perhaps too strong a word, but indeed, confidence intervals around estimates are wider under the Gamma+lethal than under the Gamma model. This comment is, we suggest, covered by our bootstrap/resampling analyses, and the fact that we consider distinct plausible values for p_{lth} throughout.

Third, as the authors say, the results for Ne depend heavily on the very strong assumption that s_bar has the same value for one set of genes in Drosophila and a different set of genes in primates. Chen et al. 2017 and Loewe et al. 2006 present methods for estimating the strength of selection directly. Could these be used to test this strong assumption?

Our understanding of the methods of Chen et al (2017) and Loewe et al (2006) is that they are similar in spirit with the method we are developing here, although the models differ in their details. Chen et al. indeed discuss the among species difference in DFE shape, as we recall (p17, l497-499). We now cite Loewe et al (2006), who like us assumed that the species they analyzed (two species of fruit flies) shared a common DFE shape. Thanks for mentioning this.

3- Simulations

Figure 4 presents simulation runs, aiming to show that piN/piS is less affected by demographic events than pi_S. This is an important part of the paper, but I was not sure how well the simulations related to the results reported.

First, the S estimates used the full SFS for non-synonymous and synonymous polymorphisms, while the simulations report piN/piS. These quantities might equilibrate differently (as is evident from Tajima's D).

We considered π_N/π_S instead of S because this is a widely used statistics, so we thought many readers would be interested in these results. We modified Figure 4 such that the response of the estimated S, not π_N/π_S, is shown. Similarly to our main result, this figure shows that the estimated S equilibrates faster than π_S.

Second, if I have understood correctly, the "recovery phase" shows a gain in genetic diversity starting from a large but genetically uniform population. Is this realistic? Why did the authors not explicitly simulate the recovery from a bottleneck?

Just for the sake of simplicity. Our procedure is equivalent to simulating recovery from a very strong bottleneck.

Third, to place their results in context, the authors cite Brandvain and Wright 2016 regarding the different equilibration times of piN and piS. But I think the explanation of the simulation results might be that ratios of pairwise diversity equilibrate more rapidly than raw diversity measurements. This is seen in neutral Fst, for example (Pannell and Charlesworth 2002).

This is a good point but since we now also show the results for the estimated S, which is not a ratio (or not obviously so), we found it difficult to include this in the manuscript.

Finally, if the authors suspect that non-equilibrium effects explain some of their results, why do they not test for these effects directly in their data (e.g. by reporting the r_i from Galtier 2016, or Tajima's D for synonymous sites etc.)?

We now report the estimated r_i's from our main analysis (new figure S7).

Minor suggestions:

The gamma shape parameter is defined in different ways, so it might be helpful to include an equation.

Although there indeed exist different parametrizations for the Gamma distribution, all of these share the same shape parameter. They rather differ in how the scale is characterized. We here parametrize the scale by specifying the mean of the distribution, following Welch et al. (2008), which we think is unambiguous.

The simulation methods state that the negative gamma distribution has a mean of -2.5. Does this apply to 4Ns, and if so, what happens when N changes?

The relevant parameters in this simulations are $\theta=4N\mu$, $\rho=4Nr$, and $S=4Ns$. These were set to 0.88 10^{-2}, 4 10^{-3} and 10^{5}, respectively, in our basic condition (before bottleneck), and twice higher in our other condition. Any set of N, μ, r and s consistent with these values are equivalent. In order to minimize computation time, and following a suggestion of the SLIM manual, we used

relatively low values for Ne, and relatively high values for μ, r and s – hence the -2.5 figure for the average selection coefficient. We could equivalently have set $N=10^6$ and mean s = -0.025. We added a sentence explaining this trick in the Material and Methods (p9, l249-251).

data sets with few polymorphisms, and all multi-allelic sites were excluded. Are the authors confident that this does not introduce biases?

We did not specifically analyze the effect of multi-allelic sites in this study, but this has been checked in previous analyses of similar data sets (e.g. in Romiguier et al 2014) and shown to have a negligible impact.

Reviewer 2

This is an interesting paper where the authors investigate Ne across species, and are lead to modify the dfe. Indeed they plausibly motivate an extension of the gamma form of the dfe, to include an effectively lethal class. While the paper is quite nicely written, some of the introductory material could be made more accessible, so a wider set of readers can benefit from this paper. Thus in particular, I make the suggestion that the following points are addressed.

1. P3 Different Ne ’s are mentioned in the literature. could the authors explicitly state the one they mean

In the introduction we are thinking in terms of a Wright-Fisher population, in which Ne is unambiguously defined.

2. P3 The authors say "In a Wright-Fisher population the power of drift is inversely proportional to the effective population size, Ne, ..." but what is precisely meant by "the power of drift"?

We now clarify that the power of drift is the across-generation variance in allele frequency due to random sampling of organisms (p3, l70-71).

3. π is mentioned and results alluded to. It would be helpful to provide some known concrete results for this, and its implied connection with heterozygosity

In the introduction we use the terms "genetic diversity" and "heterozygosity" as synonyms since they are equivalent in a panmictic, Wright-Fisher population.

4. P3 Expand a bit on Lewontin’s paradox

The text is now more explicit about what Lewontin's paradox is, mentioning census size (p3, l80-81).

5. P4: The authors say " Secondly, π_N/π_S is expected to approach its equilibrium faster than when Ne fluctuates". When Ne fluctuates could the authors say what is known about π_N/π_S and what equilibrium means, in this case. What sort of fluctuations of Ne are envisaged or the results restricted to?

We rephrased: "when Ne fluctuates" → "after a change in Ne".

6. P4 They also say "For this reason π_N/π_S might be less sensitive than π to ancient bottlenecks and selective sweeps. Empirically, there is evidence that π_N/π_S is negatively correlated to population size in Drosophila..." These statements are a bit worrying. They suggest the statistic of choice by these authors is not well understood in the sort of scenarios considered (the use of might, the appeal to empirical results). Is this really the case?

The literature on this topic – response of π_N/π_S to variation in Ne – is indeed a bit scarce; still we agree our text is more worrying than needed. We replaced "might" with "should" and slightly rephrased the next sentence.

7. P5. "the ratio of S" means ratio of what to what, S values of different species?

Yes, now clarified.

8. Lastly, the authors talk about Ne . They do not, as far as I can see, refer to actual population sizes, yet Ne is, in some way related to the population size. Could they say if their estimates of Ne scale with the census size in a way that is plausible?

In this study we estimate the Ne.s product, not Ne per se. To compare with census size, we would need some empirical estimate of the average strength of purifying selection on non-synonymous mutations – but we don't have this.

Reviewer 3

In this study, the authors use variation in mutation load among different species to assess variation in the effective population size (Ne). Doing so, they find variation in Ne in the order of 102, which is roughly one magnitude larger than variation in Ne observed based on neutral genetic diversity. In addition, the authors raise difficulties in assessing mutation load via a gamma-distributed distribution of fitness effects (DFE). Instead, the authors suggest to use a gamma-distributed DFE after subtracting an Ne-independent fraction of lethal mutations.

I find the idea to assess variation in Ne based on mutation load relevant and interesting, but I have several major concerns regarding the robustness and presentation of the results.

Besides, note that page numbers and line numbers would have been helpful. Below, I use page numbers starting from the title page as page 1.

Major remarks:

1) It would be helpful, if the authors could clarify what the focus of the study actually is. The title and the final paragraph of the Discussion section indicate that the main topic of the study is variation in Ne among different species, which would nicely tie to a discussion about different definitions and usage of Ne. However, a huge part of the manuscript circulates around methodological limitations in assessing

mutation load via a gamma-distributed distribution of fitness effects (DFE). While the latter is an interesting observation, the huge emphasis on this topic in the Discussion section appears to distract the reader from the supposed main topic. I suggest the authors to streamline the manuscript towards in more detail discussing observed differences in variation in Ne, once assessed based on mutation load and once assessed based on neutral genetic diversity. Different definitions and/or usage of the effective population size (Ne) are mentioned in the final section of the Discussion. Given the relevance of a proper understanding of Ne for transferring the main result of the study, different definitions and usage of Ne should be properly introduced in the opening of the Introduction. For example, it seems important for the reader to be aware of differences in contemporary Ne and long-term Ne. In addition, more care should be paid to the usage of Ne throughout the manuscript. For example, page 4, "... one cannot safely assume that π varies among species proportionally to Ne ..." should rather read "... one cannot safely assume that π varies among species proportionally to the number of reproducing species ...". Otherwise, please clarify what definition of Ne is referred to?

Thanks for this comment. We understand the reviewer's feeling about the lengthy discussion on the methods and how it distracts the reader. On the other hand, the approach we're taking here is new, and our results dependent on its reliability. Reviewer 1, in contrast to this reviewer, incites us to provide even more methodological material and discussion. To account for the diversity of perspectives, we structured the discussion in five well-defined, numbered subsections – 3 methodological, 2 biological – and added a header suggesting that readers not so interested in the method should jump directly to subsections 4 and 5 (p15, first paragraph). As far as the introduction is concerned, we now clarify that we theoretically consider the Ne of a Wright-Fisher population (which is unambiguously defined, p3, end of first paragrah), and we introduce the concepts of long-term vs contemporay Ne, as suggested (p4, l85).

2) The discussion around Lewontin's paradox seems somewhat to be based on a misunderstanding. Specifically, Richard Lewontin pointed out that variation in the amount of neutral genetic diversity of different species did not reflect the large amount of variation observed in census population size of different species. Lewontin's paradox therefore concerns the discrepancy between variation in neutral genetic diversity and census population size. Lewontin's paradox does not deal with different definitions and/or usage of the effective population size (Ne). Since the authors estimate variation in Ne based on mutation load rather than neutral genetic diversity, the authors do not resolve Lewontin's paradox, but circumvent Lewontin's paradox. This is an important difference, and results should be discussed accordingly.

Thanks for this comment. We agree. We now explicitly mention census size when first introducing Lewontin's paradox (p3, l80-81). We modified the abstract (p2, l47) and discussion (p22, l643) to avoid suggesting that our study addresses Lewontin's paradox. Our work is still relevant to Lewontin's paradox, we believe, in that it contributes to narrowing the gap between genetical and ecological estimates of the among-species variance in population size.

3) I find it hard to see any systematics in the comparison of different model fits. For example, for the gamma-distributed DFE the authors test if a model allowing beta to vary among species is a significantly better fit than a model assuming a common beta. However, such test appears not to be performed for the gamma-distributed DFE after subtracting a common fraction of lethal mutations (referred to as gamma + lethal DFE). I would like to see such a test, to get statistical support if the huge variation in beta is actually resolved after subtracting a common fraction of lethal mutations. Also, what are the species-specific estimates of beta after subtracting a common fraction of lethal mutations? A figure similar to Figure S3 for the gamma + lethal DFE would be helpful. More generally, I would

like to see a more systematic approach to the model comparisons. In particular, model comparisons for the “gamma-distributed DFE” and “the gamma-distributed + lethal DFE” should follow the same systematic. It would further help if model comparisons could be summarized in a table. At the moment, the choice of different models leaves a somewhat arbitrary impression, and seems motivated by verbal arguments rather than statistical support. Perhaps I am wrong, but I would like to see this clearly demonstrated.

The reviewer is correct: the shared-shape vs one-shape-per-species test is still significant under Gamma+lethal model, as we now indicate (p11, l321-323). The difference in log-likelihood between the two parametrization schemes, however, was considerably reduced when the p_{lth} parameter was introduced.

Please note that our aim in this study is not to select the model providing the best fit to the data, or has the smaller AIC. Rather, we have the constraint of choosing a plausible model for the DFE with a common shape across species – otherwise, the variation in estimated S does not inform on the variation in Ne, as shown by figure S1. The Gamma+lethal model clearly does a better job than the Gamma model under this constraint.

4) A scatter plot showing mean πN/πS versus estimates of (Nes)^(-beta) for both models, the “gamma-distributed DFE” and “the gamma-distributed + lethal DFE” would be informative. It would be interesting to see if data confirm that mean πN/πS is proportional to (Nes)^(-beta) for both models, the “gamma-distributed DFE” and “the gamma-distributed + lethal DFE”. Deviations from such a relationship could be informative about violations of underlying model assumptions.

Thanks for this suggestion. We added the suggested figure as figure S3 in the new version, which indeed reveals the existence a linear relationship between the mean πN/πS and the estimated (Nes)^(-beta), as expected at equilibrium with Gamma-distributed effects (p12, l339-341).

5) The simulation study should be extended and moved to the Results section. In particular, I would like to see how πN/πS and πS behave after a sudden increase and a sudden decrease in population size, both starting from an equilibrium value. The latter seems to be shown (though the choice of the starting value is unclear to me), but the former appears to be cut off before the equilibrium of πN/πS is reached. At the moment, I therefore get the impression that a drop in either πN/πS or πS asymptotes faster than an increase in either of them. The evidence that πN/ πS reaches its equilibrium faster than πS seems thus not convincing to me. Also, it would be helpful if the asymptotic value would be indicated by a horizontal line.

We re-ran simulations and waited a bit longer after the bottleneck at generation 15,000 (modified figure 4). The reviewer is correct that the lag in equilibration time between the estimated S (which is now our statistics of interest here, see above response to reviewer 1) and πS is only perceptible after a sudden population increase – i.e., posterior to a short-lived bottleneck. This is what we had in mind when discussing this phenomenon: ancient bottlenecks might affect diversity more than mutation load. We more specifically refer to ancient bottlenecks when discussing this point in the revised version (p20, l598 and 601).

Minor remarks:

1) The opening of the Introduction uses some awkward formulations. For example, “hazards of reproduction” suggests that there is a risk in reproduction rather than randomness. Besides, genetic drift represents the biological interpretation of stochasticity or variance in allele frequencies. It is therefore not clear to me how the variance in allele frequencies can influence the probability of fixation? Other major evolutionary forces that impact changes in allele frequencies should be mentioned alongside the introduction of genetic drift.

"hazard" was replaced with "randomness". The variance in allele frequency does not influence the fixation probability of a neutral allele, which is equal to its current population frequency irrespective of Ne, but it does affect the probability of fixation of selected alleles, as highlighted by, e.g., Ohta – more variance decreases the chances that the advantageous allele eventually "wins".

2) Page 9, Results section “Gamma DFE”, states “The optimal fruit fly beta was firmly rejected by primates as a group, and reciprocally.” Does the test include D. santomea?

Yes it does, as we now clarify (p10, l279).

3) Axis labels should be added to Figures S4.

We now clarify in the legend to figure S4 what the axes are.

4) Throughout the text, “Gamma DFE” should better read “gamma-distributed DFE.

"DFE" stands for "distribution of fitness effects", so "gamma-distributed DFE" seems to be repetitive?

5) Page 2, abstract, the authors state “..., which implies to jointly infer the DFE.” It is unclear to me what is jointly estimated here? Jointly among species?

Jointly with the magnitude of variation in drift power among species. "jointly" was replaced with "also" to remove the ambiguity

6) Page 2, abstract, “allele frequency distributions” should read “allele frequency spectra”.

Modified as suggested.